Institutional vs Generic Web Search: A Results-Centric Comparison

Baden Hughes [HREF1], Research Fellow, Department of Computer Science and Software Engineering [HREF2] , The University of Melbourne [HREF3], Victoria, 3010, Australia. badenh@csse.unimelb.edu.au

Margaret Ruwoldt, Acting Manager - Web Services [HREF4], Information Services [HREF5] , The University of Melbourne [HREF3], Victoria, 3010, Australia. m.ruwoldt@unimelb.edu.au

Abstract

We conduct a principled comparison of search results found using an institutional web search engine, and those found using a generic web search engine. We use the top 100 searches recorded against the institutional web search engine during 2005, and programatically execute the same queries against the institutional web search engine, and a selection of broad coverage web search engines, collecting the top 1000 ranked results. These results are compared across various dimensions as a method of evaluating the overall performance of the institutional web search engine across a number of metrics. The motivation for this work is twofold: both to establish a methodology for benchmarking institutional web search engine performance, and to acquire baseline measurements in the light of a pending infrastructure upgrade.

Introduction

Search engines in many contexts have become a default information source for many users. Given the size of the internet, effective use of search engines is the only viable way for end users to locate materials of interest to their information needs. It is clear that web search is a market which is open for competition, there are literally thousands of web search engines which can be selected by a user in pursuit of information. However, it is well known that search engine coverage of the web in general is quite different from engine to engine, owing to variant crawling and ranking methodologies. While there are a number of studies which consider the nature of these differences, to date a single evaluation methodology has not been accepted by the research community - meaning that comparative evaluations of search engines are usually isolate and it is difficult to correlate studies against each other.

In contrast, in the institutional context, users are often provided with a single web search engine for corporate content. Often this search facility is impoverished compared to the functionality that a user expects based on broader web search experience. In addition, the constrained domains in which corporate or institutional web search engines are applied allow for local manipulation of results according to various critiera, often in an attempt to reduce the complexity of the information discovery process. Of course, many entities which have a public web presence are also indexed by external search engines, and a competition emerges between enterprise search and internet search applications.

In this paper we conduct a principled comparison of search results found using an institutional web search engine, and those found using a generic web search engine. We use the top 100 searches recorded against the institutional web search engine during 2005, and programatically execute the same queries against the institutional web search engine, and a selection of broad coverage web search engines, collecting the top 100 ranked results at a single point in time. These results are compared across various dimensions as a method of evaluating the overall performance of the institutional web search engine across a number of metrics. The motivation for this work is twofold: both to establish a methodology for benchmarking institutional web search engine performance, and to acquire baseline measurements in the light of a pending infrastructure upgrade.

The baseline measurements have two additional practical dimensions. First, we wish to understand in greater detail the differences between search result sets provided by the institutional search engine and the search result sets provided by a range of third party broad coverage web search engines for the same query. Second, we seek to establish a baseline methodology and quantitative measures against which we can conduct future evaluations, both for purposes of measuring the effectiveness of the institutional search engine and for web meta-content modifications (navigational structures etc).

The structure of this paper is as follows. First we review some related work in the comparison of search engine results. Next, we describe our methodology in detail, after which we present a range of results from our experiments. We review and evaluate our findings, and propose a number of directions for future work. We conclude with some observations about the overall approach and results, and the likely impact.

Related Work

The idea of results-centric comparison of web search engines has been considered by a number of researchers. Much of this work has been in the area of methodologies and algorithms for the calculation of similarity between search engine result sets for given queries.

Losee and Paris (1999) measured search engine performance in terms of result sets in comparison to the perceived complexity of queries themselves. Our methodology is similar in terms of the result set comparison (we adopt URI set wise and URI rank wise similarity measures), but different in that we are not considering the relationship between the result variance and properties of the query (eg length, syntactic or semantic ambigutity) per se. Haveliwala et al (2002) demonstrated the use of an automated evaluation mechanism to replace user feedback. Their mechanism was based on a more complex result set similiarity measure than ours, primarily because of the multi-dimensional evaluation of interest in their context. Cafarella and Etzioni (2005) experimented with a range of methods for search engine comparison from the perspective of raw performance. Our measurement methodology is similar in terms of the URI set and rank wise baselines, but different in the sense that we care little about the raw speed of any given query.

At a practical level, search engine comparison is a commercially viable service offered by a range of providers in the "search engine optimization" space. There are myriad online tools (both free and pay-for-use) which allow comparisons between search engines, primarily on a per URI results rank basis. While these types of services are arguably valuable for business intelligence purposes, they lack the broad coverage evaluation that we seek: essentially these types of services offer a per URL evaluation which is not the primary target of our investigation.

In this paper, our innovations are both to compare an institutional web search engine against external broad coverage web search engines, and to develop a framework by which such a comparison can be drawn on an ongoing basis.

Methodology

The basic design principle for our work is to compare the search results for a given number of queries between the incumbent institutional search engine and broad coverage web search engines known to be widely used by local users. In order to conduct this experiment, a number of steps are required.

First we have mined the institutional web search engine log to find the most popular queries during the 2005 calendar year. Care is taken to remove from the logs searches which are not conducted by humans (eg through removing clearly identified crawler generated queries), and those which are not real searches but user errors in submitting empty or pre-populated web search forms (again, through removing default search box text). We divide the remaining searches into several categories: the top 10, top 100 and top 1000 queries. Using these divisions we can measure various aspects at a fixed point within overall queries (aka precision at a given depth in traditional information retrieval).

By way of example, we show the top 20 queries for 2005 drawn from aggregated institutional web search logs for the calendar year:

```
employment, jobs, careers, library, career, alloc8, careers online, email, bookshop, medicine,
map, housing, psychology, scholarships, law, themis, handbook, mba, exchange, human resources
```

Within this selection a number of the queries are generic (consider: careers, career, jobs) while others are highly specific to the institution (Alloc8 is the student timetabling system; Themis is institutional enterprise resource management system with staff self service; Careers Online is a database that employers and students can post job advertisements and applications to). In all cases, there is additional context which is implicit: all queries assume the subject domain of the institution ("The University of Melbourne"), and hence should be interpreted in this way. We remove from the query sets two identifiably pre-composed queries (that of the default search example in the production service, and similar text from web templates).

Having identified the most popular queries, we then programmatically interrogate both the institutional and external search engines gathering the result sets for each query. This is an automated process, and uses publicly available APIs for executing queries. The web search engines selected for comparision are Google (Google, n.d.) [HREF6], Yahoo (Yahoo, n.d.) [HREF7] and MSN (MSN, n.d.) [HREF8]. This selection is based on web server referrer log analysis for the main University web site. The institutional web search engine is UltraSeek (UltraSeek, n.d.) [HREF9].

For the external engines, the original queries are slightly modified to ensure that searches are constrained appropriately by prefixing a domain-specific restriction to each (eg site:unimelb.edu.au in the case of a Google search). Similarly, the modification of any inline syntax in the queries is required for each target web search engine to the functional equivalent expression for a given context.

For each query, we record the top 1000 results from each search engine. These initial result sets are case folded, encoding normalised to UTF-8 and consequently reduced to include only the URI and the corresponding rank within results, and combined to form a tuple, vis: query, engine, URI, rank.

We are now in a position to compute the result set similarity for each query. We are interested in two different types of result set similarity. The first we label *URI set-wise similarity*, that is, determining how many URIs appear in both sets of results. The second we label *URI rank-wise similarity*, that is, for a given URI, how the positions of URIs within result sets vary.

To compute URI set-wise similarity, we use pointwise mutual information (Church and Hanks, 1990) as the basis of our comparison. In probability theory and, in particular, information theory, the mutual information, or transinformation, of two random variables is a quantity that measures the mutual dependence of the two variables. In our case, the variables are the result set lists from each query. Mutual information measures the information about a given list that is shared by another list. If the lists are completely unique, then list A contains no information about B and vice versa, so their mutual information is zero. If list A and list B are identical then all information conveyed by list A is shared with list B: knowing list A reveals nothing new about list B and vice versa, therefore the mutual information is the same as the information conveyed by list A (or list B) alone, namely the entropy of list A. In particular we take the top 100 results for each query for each search engine, and compute the mutual information between each list.

To compute URI rank-wise similarity, we use the Spearman rank correlation coefficient (Lehman and D'Abrera, 1998) as the basis of our comparison. Spearman's rank correlation coefficient is a non-parametric measure of correlation: that is, it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables.

In the case of Google, Yahoo and MSN, we have been provided with unrestricted access in terms of number of queries dispatched to the respective search APIs, but we are still restricted to the number of results returned per query (1000). In all cases we interact programmatically with the search engine via SOAP over the publicly available API.

The University of Melbourne uses the Ultraseek search engine, which does not come with a formal API for programmatic interaction. However, Ultraseek does have a systematically constructed URI sequence for queries, and in this experiment we wrote a small URI generator facility which generates the appropriate URI with query strings embedded, and then use a simple HTTP POST to generate the results from the search engine. Result sets are obtained by extracting the URIs from the returned HTML.

The entire experimental suite is written in Perl, using a combination of libwww, a range of search engine interaction modules, and service-oriented middleware such as SOAP. The script directly produces statistical results which can be visualized in various ways. While there are a number of alternative ways this task could be done, a core motivation in our context is automation of the process so that it can be repeatedly conducted without the need for human interaction.

Results

Having described our methodology in depth, we now turn to a detailed presentation of a selection of our results, keeping in mind that our two basic questions are essentially: How much variation is there in the results returned by different search engines? How much variation is there in the rankings returned by different search engines?

Figure 1. Frequency of Query Occurrence by Rank

In Figure 1 we show the overall shape of the frequency of occurrence of queries, this graph forms the familiar 'long tail' graph common in web evaluations of all kinds. This reinforces our selection mechanism of the top 1000 queries, and the binning approach of top 10, top 100 and top 1000 queries. While there are many queries below the top 1000, their frequency of execution is logarithmically smaller than the top 1000 queries we select in these experiments.

Figure 2. UltraSeek URI Set-Wise Similarity vs External Engines

In Figure 2 we show the URI set wise similarity between the institutional and external search engines at a given

result depth, on an individual basis of external web search engines against the institutional web search engine. We can see that in regard to the number of URLs in common between engines, that as the number of results increases, the similarity decreases, to almost 50% differences in the URLs returned between engines at a depth of 1000 results. From this analysis we can see that that the higher ranked a given query, the more likely that the result set returned by search engines is to be similar. This is to be expected since the queries increase in specificity as they decrease in ranking; that is, higher ranked queries are more generic, and lower ranked queries are more specific.

Figure 3. UltraSeek URI Rank-Wise Similarity vs External Engines

In Figure 3 we show the URI rank wise similarity between the 4 search engines at a given result depth, both on an individual basis for external web search engines against the institutional web search engine, and in aggregate. The Y axis shows the result depth, while the X axis shows the degree of similarity (100% = identical; 0% = no overlap). From this analysis we can see that as the number of queries increases, so the similarity between the result sets decreases, but certainly not in a linear fashion. Expecting that there will be some differences between the results returned by each search engine, it seems that there is empirical evidence for this variation.

Figure 4. Overall (Interpolated Set-wise and Rank-wise) Similarity Between Search Engines

In Figure 4 we show the overall similarity between all search engines based on the top 100 search results for the top 1000 queries. We interpolate set-wise and rank-wise similarity to arrive at this overall assessment. From this analysis we can see that in the highest frequency queries and the highest ranked results, there is significant variation between the search engines. This in turn implies that the picture presented to the end user in terms of available resources to fulfill their information need is quite different depending on the selection of search engine. As such, we can expect that users will draw not only a different conclusion about the available resources in answer to their enquiry, but also that simplistic (ie manual) query result comparisons (probably based on the "if all else fails try Google" methodology) highlight these differences.

Discussion and Future Work

From these results it is obvious that there is a significant difference between the search results found in The University of Melbourne's search engine and those found for the same queries in commercial competitors. While this is perhaps not unexpected, the degree of variation is such that it is obvious that users who prefer an external search engine over the institutional engine are having their information needs interpreted and fulfilled quite differently. Which engine is more effective in that context is not clear; what is now strongly evidenced is that the results are different depending on which engine is queried against.

One key difference between the view of web content seen by the institutional search engine and external search engines is that there are a number of web pages at The University of Melbourne which are restricted access, particularly by layer 4 or higher restrictions eg IP subnet. It is therefore correct to assume that the institutional web search engine has access to a different topology of institutional web content than external engines, and this may account for some of the variation. However, a manual inspection of a random set of 50 URIs returned by the institutional search engine showed only 12% of those were not externally accessible, and so we conclude that the differences based on this factor are not accounted for by this alone.

The URI similarity (both set-wise and rank-wise) improves significantly if we only include URIs of MIME type text/htm and text/html, compared to URIs of all types. Most notably, around 52% of URIs are of a different MIME type, particularly msword/doc, mspowerpoint/ppt and text/pdf. In other experiments we found that overall set-wise similarity increased between search engines if we filtered out the richer document types from the institutional search engine result sets: by 13% if the application/msword MIME type was excluded, by an additional 6% if the application/mspowerpoint MIME type was excluded and by an additional 8% if the pdf MIME type was excluded; an increase of 27%. For rank-wise similarity, an overall increase of 21% was found (doc +6%, ppt +8% and pdf +7%). What these results show us is that the similarity between search engines is in fact MIME type dependent. An immediate performance improvement in terms of increased similarity would be to remove from the Ultraseek indexes at least any Microsoft Word or PowerPoint objects, and arguably, even PDF documents.

We have not specifically attempted to weed out non-human queries from the overall set of popular queries; the search engine logs do not identify a user agent and as such any removal of automated enquries would be difficult at best. Random manual inspection of 10% the top 1000 queries does not immediately reveal any queries which

appear to be non-human in terms of their properties.

We are only computing over top 100 results per query, although we have access to 1000 results per query. This is not a computational constraint, but rather a pragmatic one: users are unlikely to progress through result sets to depth 100, let alone depth 1000. Since at depth 100 we already see significant variation, there is no cause for us to believe that the situation would improve by including more results.

Several items of future work arise from our experience.

The first is to fully automate the comparison framework for ongoing evaluation: while at this point in time the framework is automated from the perspective of analysis, it requires human intervention to provide the top 1000 query data, to instantiate the analytical cycle, and to convert the resulting statistics into a graphical display. These constraints are not atypical of the first generation of such software, and clearly if the evaluation framework is adopted as an ongoing measure of performance, then such improvements will be made.

The second is to improve the analysis framework to take into account the differences in results based purely on access control measures: a reasonable proportion of The University of Melbourne's web content is restricted by IP subnet level access controls. Such controls prevent external web search engines from indexing the web content, but do not prevent the institutional web search engine from crawling and indexing the same pages. It is unclear at this point as to the degree of impact of this access level difference: a systematic evaluation of segments of the institutional web content and its corresponding access controls is required to ensure that issues arising from this access restriction are not statistically significant.

We would also like to correlate our findings against several other data sources which at the time of writing we have been unable to access, including:

- Comparisons with other commonly used search engines, we have simply selected the top 3 external engines used based on the http referrer statistics of The University of Melbourne's main web site, but there are numerous other engines which are well attested;
- It would be useful to review the correlation of query types and frequencies against the queries executed by users via the outbound proxy at The University of Melbourne, we are making the assumption that the query properties are similar in both contexts.
- Also we wish to consider end user click through behaviour in result sets, we have not considered whether or not there is a relation between the rank position of a given URI in a result set and its likelihood of being selected by an end user for a given query.

Finally, the research described here is based on statistical comparisons of data sets. It does not address web user's subjective experience of using the various search engines. Qualitative usability research, in combination with the regular empirical performance data, would enable an institution to identify areas of its web search service and overall web site that may require changes to content, structure and possibly enterprise information architectures. We expect this research and redevelopment to become an iterative cycle as web users' behaviour and expectations change over time.

Conclusion

We have described and demonstrated a method for principled comparison of institutional and broad coverage web search engines. Our findings are that the incumbent institutional web search engine at The University of Melbourne provides substantially different results from broad coverage web search engines such as Google, Yahoo and MSN for a given set of popular queries, in both URI occurrence and URI rank similarity dimensions. A range of short term improvements can be recommended based on an analysis of the differences between the institional and broad coverage web search engines; the most effective is the pruning of the institutional web search engine indices to exclude any non-HTML document types, which has been done since our original study. Our contribution in this paper is to develop an open, repeatable process by which such analysis can be performed on an ongoing basis, and to establish a baseline for measuring the effect of any future institutional web search engine enhancements.

References

Taher H. Haveliwala, Aristides Gionis, Dan Klein and Piotr Indyk, 2002. Evaluating Strategies for Similarity Search on the Web. Proceedings of the 11th International World Wide Web Conference (WWW2002). Association for Computing Machinery. pp.432-442.

Michael J. Cafarella and Oren Etzioni, 2005. A Search Engine for Natural Language Applications. Proceedings of the

14th International World Wide Web Conference (WWW2005). Association for Computing Machinery. pp.442-452.

Robert M. Losee and Lee Anne H. Paris, 1999. Measuring Search Engine Quality and Query Difficulty: Ranking with Target and Freestyle. Journal of the American Society for Information Science 50(10), pp.882-889.

E. L. Lehman and H.J.M. D'Abrera, 1998. Nonparametrics: Statistical Methods based on Ranks. Prentice Hall: New Jersey.

Kenneth W. Church and P. Hanks, 1990. Word association norms, mutual information, and lexicography. Computational Lingustics 16, pp.22-29.

Google Inc., n.d. Google Search Engine. http://www.google.com. Available online at [HREF6]. Last Accessed 15 March 2006.

Yahoo Inc., n.d. Yahoo Search Engine. http://www.yahoo.com. Available online at [HREF7]. Last Accessed 15 March 2006.

Microsoft Inc., n.d. MSN Search. http://search.msn.com. Available online at [HREF8]. Last Accessed 15 March 2006.

Ultraseek, n.d. UltraSeek, The University of Melbourne's Web Search Engine. http://websearch.its.unimelb.edu.au. Available online at [HREF9]. Last Accessed 15 March 2006.

Hypertext References

HREF1
http://www.csse.unimelb.edu.au/~badenh/
HREF2
http://www.csse.unimelb.edu.au/
HREF3
http://www.unimelb.edu.au/
HREF4
http://www.unimelb.edu.au/webcentre
HREF5
http://www.infodiv.unimelb.edu.au/
HREF6
http://www.google.com/
HREF7
http://www.yahoo.com/
HREF8
http://search.msn.com
HREF9
http://websearch.its.unimelb.edu.au/

Copyright